# Exploring Photorealistic Real-time Image Synthesis

Haifeng Huang

## 1. Introduction

Generative Adversarial Networks (GANs) are capable of generating high-quality images; however, the resolution of generated images remains relatively small. There were many efforts to address this issue. For example, ProGAN trains high-resolution GANs in the single-class setting by iteratively training across a set of increasing resolutions. Nevertheless, the model training is still unstable regardless of the large number of studies that have investigated and proposed improvements. Without auxiliary stabilization techniques, this training procedure is notoriously brittle, requiring finely-tuned hyperparameters and architectural choices to work at all. Most of the improvements have been made due to changes the objective function or constraining the discriminator model during the training. More recently, scaling up GAN models has been found to work pretty well for generating both high-quality and larger images.

The authors provide class information to the Generator with class-conditional BatchNorm, as seen in the image (sub-figure (a) and (b)) above, and to the Discriminator with projection. They also use Orthogonal Inialization instead of classic Xavier Initialization or N(0, 0.02I). BatchNorm Statistics in G are computed across all devices instead of per-devices, which is a typical scenario. They note that progressive growing, as ProGAN, is unnecessary. Simply by increasing the batch size by a factor of 8 improved their performance, in terms of Inception Score (IS), by 46%. They explain it that it provides better gradients for both networks. Also, they achieved a better final performance in fewer iterations. They then increase number of channels, in CNNs, in each layer by 50% (meaning the number of parameters are almost doubled). It resulted in 21% improvement in terms of IS. Notice from the figure above that class embeddings are shared and they use separate linear layers to fit each BatchNorm layer. It reduces computation cost a lot and improves training speed by 37Notice the noise vector z is split into one chunk per ResBlock and conctaenated with class embedding c. It gave a slight improvement of 4 Also, if you wonder what Non-local block is, here's is the diagram

## 2. Related Work

In the earlier I2I works [24], researchers used many aligned image pairs as the source domain and target domain to obtain the translation model that translates the source images to the desired target images. **Unsupervised I2I** Training supervised translation is not very practical because of the difficulty and high cost of acquiring these large, paired training data in many tasks. Taking photo-to-painting translation as an example (e.g., f. in Fig, it is almost impossible to collect massive amounts of labeled paintings that match the input landscapes. Hence, unsupervised methods [76, 27, 63] have gradually attracted more attention. In an unsupervised learning setting, I2I methods use two large but unpaired sets of training images to convert images between representations. **Semi-supervised I2I** In some special scenarios, we still need a little expensive human labeling or expert guidance, as well as abundant unlabeled data, such as those of old movie restoration [43] or genomics [52]. Therefore, researchers consider introducing semi-supervised learning [28, 48, 5] into I2I to further promote the performance of image translation. Semi-supervised I2I approaches leverage only source images alongside a few source-target aligned image pairs for training but can achieve more promoted translated results than their unsupervised counterpart. **Few-shot I2I** Nonetheless, several problems remain regarding translation using a supervised, unsupervised or semi-supervised I2I method with extremely limited data. In contrast, humans can learn from only one or limited exemplars to achieve remarkable learning results. As noted by meta-learning [73, 57] and few-shot learning [53, 58], humans can effectively use prior experiences and knowledge when learning new tasks, while artificial learners usually severely overfit without the necessary prior knowledge. Inspired by the human learning strategy, few- and one-shot I2I algorithms [38, 34, 35, 36] have been proposed to translate from very few (or even one) in the limit unpaired training examples of the source and target domains.

Although learning settings may differ, most of these I2I techniques tend to learn a deterministic one-to-one mapping and only generate single-modal output, as shown in Fig.. However, in practice, the two-domain I2I is inherently ambiguous, as one input image may correspond to

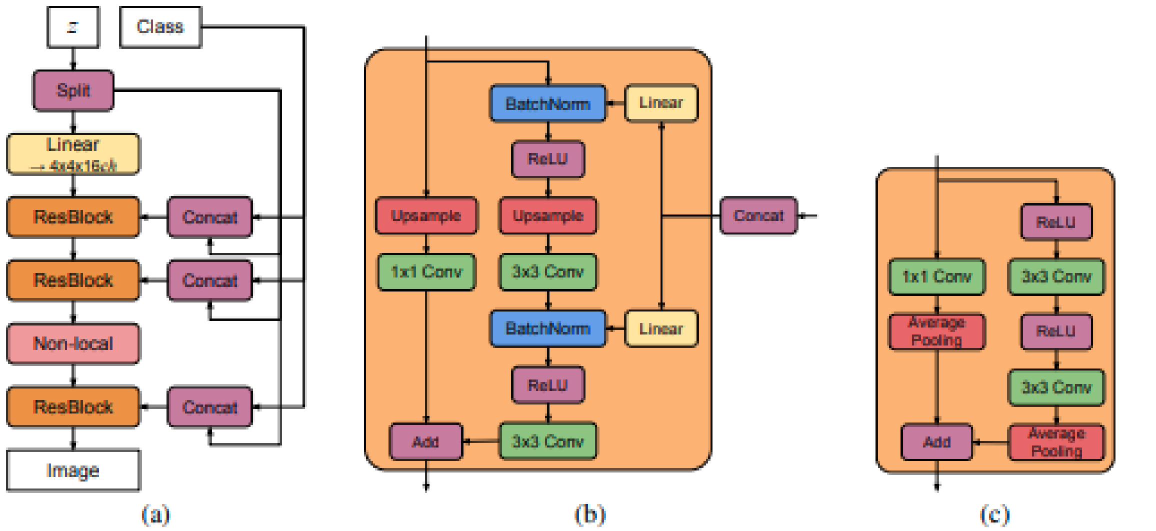


Figure 15: (a) A typical architectural layout for BigGAN's **G**; details are in the following tables. (b) A Residual Block (*ResBlock up*) in BigGAN's **G**. (c) A Residual Block (*ResBlock down*) in BigGAN's **D**.

Figure 1.

multiple possible outputs, namely, multimodal outputs, as shown in Fig.. Multimodal I2I translates the input image from one domain to a distribution of potential outputs in the target domain while remaining faithful to the input. These diverse outputs represent different color or style texture themes (i.e., multimodal) but still preserve the similar semantic content as the input source image. Therefore, we actually view multimodal I2I as a special two-domain I2I and discuss it in supervised and unsupervised settings (subsection).

Most of computer visions problems can be seen as an image-to-image translation problem, mapping an image from one domain to another image in different domain. As an illustration, super-resolution can be viewed as a concern of mapping a low-resolution image to a similar high-resolution one; image colorization is a problem of mapping a gray-scale image to a corresponding color one. The problem can be investigated in supervised and unsupervised learning methods. In the supervised approaches, paired of images in various domains are available [24]. In the unsupervised models, only two separated sets of images are available in which one composed of images in one domain and the other composed of different domain images—there is no paired samples representing how an image can possibly translated to a corresponding image in different domain. For lack of corresponding images, the unsupervised image-to-image translation problem is considered more difficult, but it is more feasible because training data collection is easier.

When assessing the image translation problem from a likelihood viewpoint, the main challenge is to learn a mutual distribution of images in different domains. In the unsupervised setting, the two sets composed of images from two minor distributions of different domains, and the task is to gather the cooperative distribution by utilizing these images. However, driving the joint distribution from the minor distributions is extremely ill-posed problem. In this section, we discuss the image-to-image translation methods. Image-to-image translation is similar to style transfer, which as the input receives a style image and a content image. The model output is an image that has the content of the content image and the style of the style image. It is not only transferring the images' styles, but also manipulates features of objects. This section lists several models that are proposed for image-to-image translation from supervised methods to unsupervised ones. Figure shows sample generate results by [24].

## 2.1. Supervised Translation

Isola et al. [24] proposed to merge the different network losses of Adversarial Network with $L_1$ regularization loss, therefore the particular generator not only trained to pass the discriminator filtering but also to produce images that contain realistic objects and similar to the ground-truth images. $L_1$ generates less blurry images as compared to $L_2$, it was the reason for using $L_1$. The conditional GAN loss is

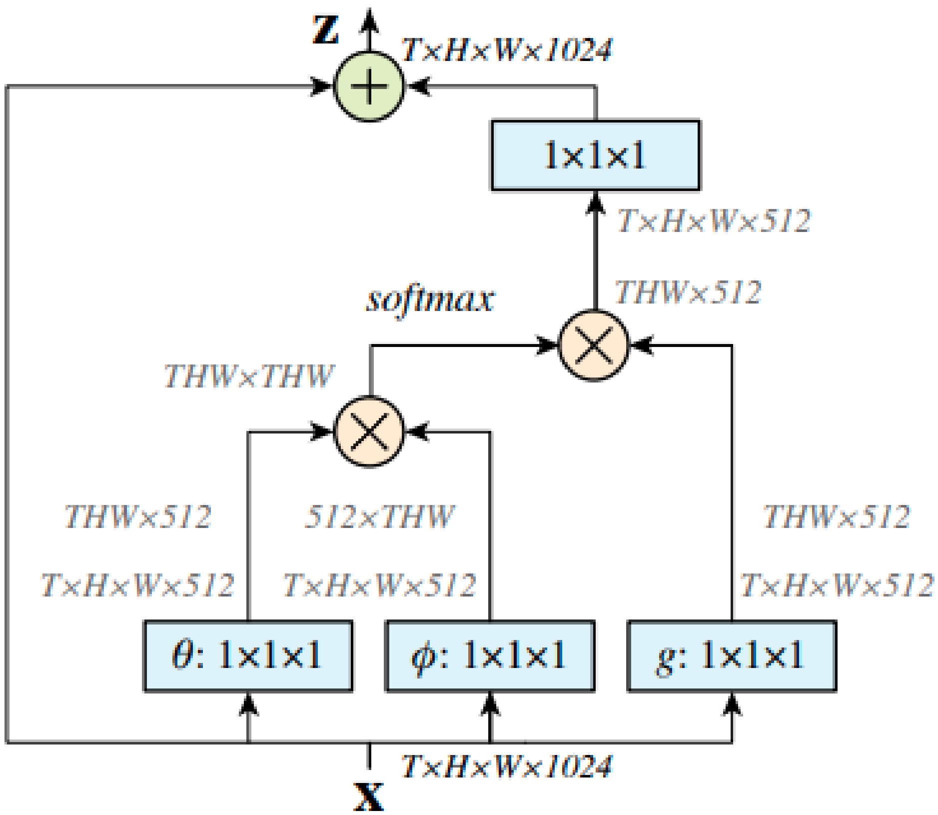


Figure 2. A spacetime **non-local block**. The feature maps are shown as the shape of their tensors, *e.g.*, $T\times H\times W\times 1024$ for 1024 channels (proper reshaping is performed when noted). "$\otimes$" denotes matrix multiplication, and "$\oplus$" denotes element-wise sum. The softmax operation is performed on each row. The blue boxes denote $1\times1\times1$ convolutions. Here we show the embedded Gaussian version, with a bottleneck of 512 channels. The vanilla Gaussian version can be done by removing $\theta$ and $\phi$, and the dot-product version can be done by replacing softmax with scaling by $1/N$.

Figure 2.

formulated as:

$$\begin{aligned}\ell_{cGAN}(G,D) = E_{(x,y)\sim p_{data}(x,y)}[\log D(x,y)]+ \\ E_{x\sim p_{data}(x),z\sim p_z(z)}[\log(1-D(x,G(x,z))].\end{aligned} \quad (1)$$

in which $x,y\sim p(x,y)$ denotes to the images that have different styles but belong to the same scene, similar to the standard GAN [18], $z\sim p(z)$ represents random noise, thereby $L_1$ loss for pressuring self-similarity is defined as:

$$\ell_{L_1}(G) = E_{x,y\sim p_{data}(x,y)}, z\sim p_z(z), [||y-G(x,z)||_1], \quad (2)$$

the general objective is specified by:

$$G^*, D^* = arg^{\min_G \max_D} \ell_{cGAN}(G,D) + \lambda \ell_{L_1}(G) \quad (3)$$

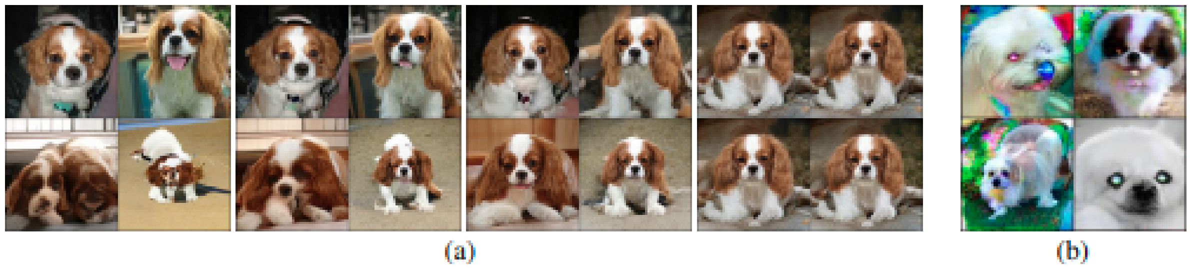


Figure 2: (a) The effects of increasing truncation. From left to right, the threshold is set to 2, 1, 0.5, 0.04. (b) Saturation artifacts from applying truncation to a poorly conditioned model.

Figure 3.

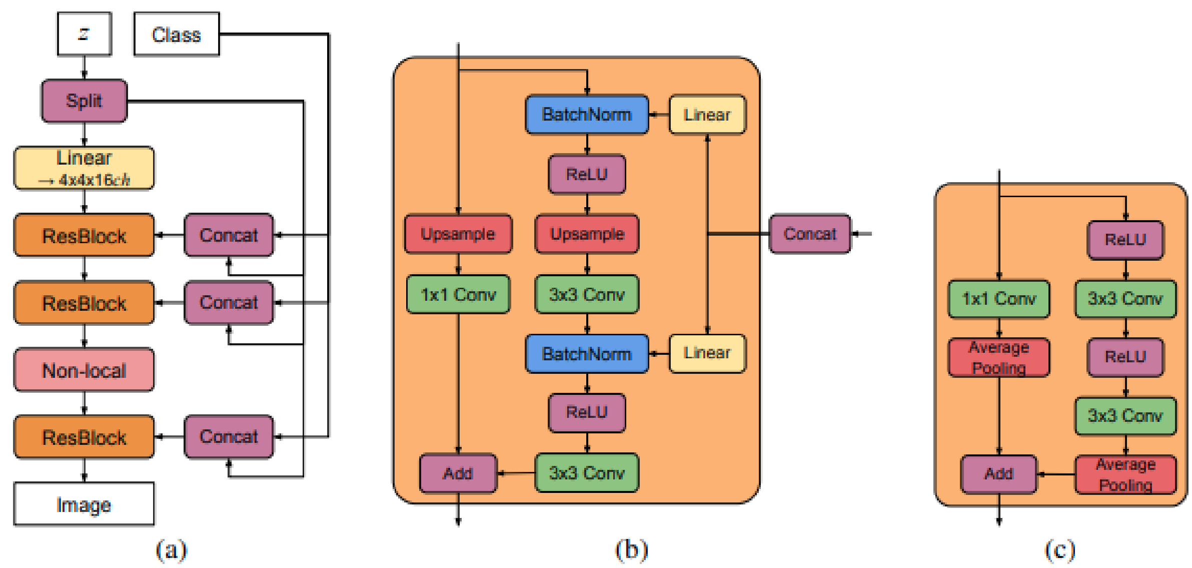


Figure 15: (a) A typical architectural layout for BigGAN's **G**; details are in the following tables. (b) A Residual Block (*ResBlock up*) in BigGAN's **G**. (c) A Residual Block (*ResBlock down*) in BigGAN's **D**.

Figure 4.

in which the hyperparameter of $\lambda$ is used to balance the two loss functions. Moreover, in [24], the authors pointed out that, the noise $z$ does not have noticeable influence on the result, therefore, they proposed to use the noise in the form of dropout during training and test in place of samples that belongs to random distribution. In this model, the structure of the $G$ is based on the new structure of U-Net that has multi-scale connections to join each encoder layer to the same layer decoder for sharing low-level information like edges of objects. In [24] the authors proposed `PatchGAN`. The proposed model rather than classifying the whole image attempts to classify the `N x N` path of each image and seek the average scores of patches for obtaining the final score of the image. From the experiments it has been observed, for obtaining the high frequency details, it is sufficient to limit the discriminator to focus on the local patches.

Yoo et al. proposed an algorithm for supervised image-to-image translation, while having a secondary discriminator $D_{pair}$ that evaluates whether or not a pair of images from multiple domains is related with each other. The loss of

$D_{pair}$ is calculated as follows:

$$\begin{aligned} \ell_{pair} &= -t\log[D_{pair}(X_s, X)] \\ &+(t-1)\log[1-D_{pair}(X_s, X)], \\ s.t.t &= \begin{cases} 0 & if X = X_t \\ 0 & if X = \hat{X}_t \\ 0 & if X = X_{\bar{t}} \end{cases} \end{aligned} \tag{4}$$

where the input image from the source domain is represented by $X_s$ and its groundtruth image is denoted by $X_t$ in the target domain, an irrelevant image in the target domain is represented by $X_{\bar{t}}$. The generator in the proposed model transfers $X_s$ into a single image $\hat{X}_t$ in the associated domain. The authors proposed an efficient pyramid adversarial networks to generating synthetic labels based on target domains for road segmentation in remote sensing images. Zareapoor et al. proposed a semi-supervised adversarial networks for dataset balancing in mechanical devices. The authors integrate multi-instance learning into adversarial networks for human pose estimation. As the results show, the proposed model has high accuracy and fast performance. Shamsolmoali et al. to handle the imbalanced class problems, proposed a capsule adversarial networks based on minority class augmentation.

In, the authors proposed a general learning framework assign the generated samples to a distribution over a set of labels instead of a single label. The effectiveness of their proposed model is proved through a set of experiments. Zhang et al. proposed DRCW-ASEG method in order to generate synthetic examples for multi-class imbalanced problem. The authors shown that their proposed strategy is able to improve the classification accuracy.

there is no noise input in the generator of pix2pix. A novelty of pix2pix is that the generator of pix2pix learns a mapping from an observed image $y$ to output image $G(y)$, for example, from a grayscale image to a color image. As a follow-up to pix2pix, pix2pixHD [61] used cGANs and feature matching loss for high-resolution image synthesis and semantic manipulation. With the discriminators, the learning problem is a multi-task learning problem. Chrysos et al. [8] proposed robust cGANs. Thekumparampil et al. [60] discussed the robustness of conditional GANs to noisy labels. Conditional CycleGAN [39] uses cGANs with cyclic consistency. Mode seeking GANs (MSGANs) [40] proposes a simple yet effective regularization term to address the mode collapse issue for cGANs. GANs are also utilized to achieve image composition [33, 3, 69, 65], Based on cGANs, we can generate samples conditioning on class labels [45, 44], text [49, 22, 71]. In [71, 70], text to photo-realistic image synthesis is conducted with stacked generative adversarial networks (SGAN) [23]. cGANs have been used for convolutional face generation [15], face aging [1], multi-modal image translation [59, 75, 67], panoramic image generation [14, 54], exemplar-based image synthesis [75, 72], synthesizing outdoor images having specific scenery attributes [25], natural image description [9], and scene manipulation [62]. Most cGANs based methods [11, 47, 51, 13, 55] feed conditional information $y$ into the discriminator by simply concatenating (embedded) $y$ to the input or to the feature vector at some middle layer. cGANs with projection discriminator [41] adopts an inner product between the condition vector $y$ and the feature vector. Two-domain I2I can solve many problems in computer vision, computer graphics and image processing, such as image style transfer (f.) [76, 31], bounding box and keypoints [50, 68] which can be used in photo editor apps to promote user experience and semantic segmentation (c.) [46, 78], which benefits the autonomous driving and image colorization (d.) [56, 32], and domain adaptation [42, 6, 37, 66].. If low-resolution images are taken as the source domain and high-resolution images are taken as the target domain, we can naturally achieve image super-resolution through I2I (e.) [64, 74].

#### 2.1.1 Multimodal Outputs

As shown in Fig.1, multimodal I2I translates the input image from one domain to a distribution of potential outputs in the target domain while remaining faithful to the input.

Actually, this multimodal translation benefits from the solutions of *mode collapse problem* [17, 2, 19], in which the generator tends to learn to map different input samples to the same output. Thus, many multimodal I2I methods [77, 4] focus on solving the mode collapse problem to lead to diverse outputs naturally. BicycleGAN [77] became the first supervised multimodal I2I work by combining cVAE-GAN [21, 29, 30] and cLR-GAN [7, 12, 13] to systematically study a family of solutions to the mode collapse problem and generate diverse and realistic outputs.

Similarly, Bansal et al. [4] proposed PixelNN to achieve multimodal and controllable translated results in I2I. They proposed a nearest-neighbor (NN) approach combining pixelwise matching to translate the incomplete, conditioned input to multiple outputs and allow a user to control the translation through on-the-fly editing of the exemplar set.

Another solution for producing diverse outputs is to use *disentangled representation* [7, 20, 26, 10] which aims to break down, or disentangle, each feature into narrowly defined variables and encodes them as separate dimensions. When combining it with I2I, researchers disentangle the representation of the source and target domains into two parts: domain-invariant features *content*, which are preserved during the translation, and domain-specific features *style*, which are changed during the translation. In other words, I2I aims to transfer images from the source domain to the target domain by preserving *content* while replacing

*style*. Therefore, one can achieve multimodal outputs by randomly choosing the *style* features that are often regularized to be drawn from a prior Gaussian distribution $N(0, 1)$. Gonzalez-Garcia et al. [16] disentangled the representation of two domains into three parts: the *shared* part containing common information of both domains, and two *exclusive* parts that only represent those factors of variation that are particular to each domain. In addition to the bi-directional multimodal translation and retrieval of similar images across domains, they can also transfer a domain-specific transfer and interpolation across two domains.

## 3. Conclusion

We find out that taking models trained with z N(0, I) and sampling from a truncated normal boosts IS and FID. Truncation trick: truncating a z vector by resampling the values having a magnitude greater than a chosen threshold. It leads to a better quality images in the cost of overall sample variety. The smaller the threshold, the smaller sample variety. where W is a weight matrix and beta is a hyperparameter set to 1e-4. They notice some of their larger models do not benefit from truncation trick. Therefore, they introduce Orthogonal Regularization due to which 60% of larger models became amenable to truncation. So, this wraps up our discussion of GauGAN's architecture and it's objective functions. In the next part, we talk about how GauGAN is trained and how does it fare as compared to it's rival algorithms, especially it's predecessor Pix2PixHD. Till then, you can checkout the GauGAN web demo, which allows you to create random landscapes. We see that the noise vector z is first split into equal size chunks. First, we take the very first chunk (zs[0]) as input and the rest chunks are used for concatenation with our class conditional vector y. After that we iterate over our ResBlock (self.blocks), as well as concatenated vectors, and pass our parameters. The final output is obtained by passing through batchnorm-relu-conv and tanh. Looks pretty simple, right? Now let's see what happens inside our BatchNorm blocks. We see that our concatenated vector y is passed into self.gain and self.bias which are just Linear layers. So, vector y is linearly projected to produce per-sample gains and biases for the BatchNorm layers of the block. The bias projections are zero-centered, while the gain projections are centered at 1. Therefore, we add 1 after we apply self.gain. Finally, after we normalize our input x, we multiply it by our computed gain and add bias. Some Last Words I hope I help someone understand the concepts of BigGAN better. Anyways, my articles are just to introduce you to the concepts. You can always read the paper and, of course, get more details from it. I encourage to study the paper on your own. This article provides a great amount of information so you the paper seem a little bit easier.

## References


[1] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017. 5

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 5

[3] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning image-conditional binary composition. *International Journal of Computer Vision*, 128(10):2570–2585, 2020. 5

[4] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. Pixelnn: Example-based image synthesis. In *International Conference on Learning Representations*, 2018. 5

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019. 1

[6] Jinming Cao, Oren Katzir, Peng Jiang, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. Dida: Disentangled synthesis for domain adaptation, 2018. 5

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Neural Information Processing Systems*, pages 2172–2180, 2016. 5

[8] Grigorios G Chrysos, Jean Kossaifi, and Stefanos Zafeiriou. Robust conditional generative adversarial networks. *arXiv preprint arXiv:1805.08657*, 2018. 5

[9] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *IEEE International Conference on Computer Vision*, pages 2970–2979, 2017. 5

[10] Emily L Denton and vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4414–4423. Curran Associates, Inc., 2017. 5

[11] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using laplacian pyramid of adversarial networks. In *Neural Information Processing Systems*, pages 1486–1494, 2015. 5

[12] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 5

[13] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016. 5

[14] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumina-

tion from a single image. *arXiv preprint arXiv:1704.00090*, 2017. 5

[15] Jon Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014. 5

[16] Abel Gonzalez-Garcia, Joost Van De Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *Advances in neural information processing systems*, pages 1287–1298, 2018. 6

[17] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks, 2017. 5

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Neural Information Processing Systems*, pages 2672–2680, 2014. 3

[19] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Neural Information Processing Systems*, pages 5767–5777, 2017. 5

[20] I. Higgins, Loïc Matthey, A. Pal, Christopher P. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 5

[21] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 5

[22] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7986–7994, 2018. 5

[23] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5077–5086, 2017. 5

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 1, 2, 4

[25] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 5

[26] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 5

[27] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865, 2017. 1

[28] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014. 1

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *stat*, 1050:1, 2014. 5

[30] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016. 5

[31] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, pages 1–16, 2020. 5

[32] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5

[33] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018. 5

[34] Jianxin Lin, Yingxue Pang, Yingce Xia, Zhibo Chen, and Jiebo Luo. Tuigan: Learning versatile image-to-image translation with two unpaired images. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020. 1

[35] Jianxin Lin, Yijun Wang, Tianyu He, and Zhibo Chen. Learning to transfer: Unsupervised meta domain translation. *arXiv preprint arXiv:1906.00181*, 2019. 1

[36] Jianxin Lin, Yingce Xia, Sen Liu, Tao Qin, and Zhibo Chen. Zstgan: An adversarial approach for unsupervised zero-shot image-to-image translation. *arXiv preprint arXiv:1906.00184*, 2019. 1

[37] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *Advances in neural information processing systems*, pages 2590–2599, 2018. 5

[38] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[39] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Conditional cyclegan for attribute guided face image generation. *arXiv preprint arXiv:1705.09966*, 2017. 5

[40] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019. 5

[41] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018. 5

[42] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[43] Aamir Mustafa and Rafał K. Mantiuk. Transformation consistency regularization – a semi-supervised paradigm

for image-to-image translation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 599–615, Cham, 2020. Springer International Publishing. 1

[44] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477, 2017. 5

[45] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pages 2642–2651, 2017. 5

[46] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

[47] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 5

[48] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in neural information processing systems*, pages 3546–3554, 2015. 1

[49] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1–10, 2016. 5

[50] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Neural Information Processing Systems*, pages 217–225, 2016. 5

[51] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *IEEE International Conference on Computer Vision*, pages 2830–2839, 2017. 5

[52] Mingguang Shi and Bing Zhang. Semi-supervised learning improves gene expression-based prediction of cancer recurrence. *Bioinformatics*, 27(21):3017–3023, 2011. 1

[53] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 1

[54] Shuran Song and Thomas Funkhouser. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6926, 2019. 5

[55] Kumar Sricharan, Raja Bala, Matthew Shreve, Hui Ding, Kumar Saketh, and Jin Sun. Semi-supervised conditional gans. *arXiv preprint arXiv:1708.05789*, 2017. 5

[56] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Infrared image colorization based on a triplet dcgan architecture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–23, 2017. 5

[57] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2019. 1

[58] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 1

[59] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2417–2426, 2019. 5

[60] Kiran K Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. In *Neural Information Processing Systems*, pages 10271–10282, 2018. 5

[61] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 5

[62] Shunyu Yao, Tzu Ming Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, Bill Freeman, and Josh Tenenbaum. 3d-aware scene manipulation via inverse graphics. In *Neural Information Processing Systems*, pages 1887–1898, 2018. 5

[63] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 1

[64] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 5

[65] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2059–2068, 2019. 5

[66] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–266, 2018. 5

[67] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 5

[68] Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, and Xuansong Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 5

[69] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3653–3662, 2019. 5

[70] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2019. 5

[71] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision*, pages 5907–5915, 2017. 5

[72] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 5

[73] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Advances in Neural Information Processing Systems*, pages 2365–2374, 2018. 1

[74] Yongbing Zhang, Siyuan Liu, Chao Dong, Xinfeng Zhang, and Yuan Yuan. Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution. *IEEE transactions on Image Processing*, 29:1101–1112, 2019. 5

[75] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11465–11475, 2021. 5

[76] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, pages 2223–2232, 2017. 1, 5

[77] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Neural Information Processing Systems*, pages 465–476, 2017. 5

[78] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5